REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
APRIL 2013	CONFERENCE PAPER (Post Print)	JUL 2012 – NOV 2012

4. TITLE AND SUBTITLE

ASSESSING TRUSTWORTHINESS IN COLLABORATIVE ENVIRONMENTS

5a. CONTRACT NUMBER
FA8750-12-C-0011

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
62788F

6. AUTHOR(S)

Michael Mayhew, Jeffrey Segall, Rachel Greenstadt, Michael Atighetchi

5d. PROJECT NUMBER
E2BB

5e. TASK NUMBER
BB

5f. WORK UNIT NUMBER
AC

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Raytheon BBN Technologies
10 Moulton Street
Cambridge, MA 02138

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/Information Directorate
Rome Research Site/RIEBA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RI

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TP-2013-003

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA Case Number: 88ABW-2012-4350
DATE CLEARED: 7 AUG 2012

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Collaborative environments, specifically those concerning information creation and exchange, increasingly demand notions of trust and accountability. In the absence of explicit authority, the quality of information is often unknown. Using Wikipedia edit sequences as a use case scenario, we detail experiments in the determination of community-based user and document trust. Our results show success in answering the first of many research questions: Provided a user's edit history, is a given edit to a document positively contributing to its content? We detail how the ability to answer this question provides a preliminary framework towards a better model for collaborative trust and discuss subsequent areas of research necessary to broaden its utility and scope.

15. SUBJECT TERMS
Collaborative Trust, Cyber Analytics, Wikipedia, Information Flow Controls

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			MICHAEL J. MAYHEW
U	U	U	UU	5	19b. TELEPHONE NUMBER (*Include area code*) N/A

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Assessing Trustworthiness in Collaborative Environments

Jeffrey Segall
Drexel University Department of Computer Science
3141 Chestnut Street
Philadelphia, PA 19104
js572@drexel.edu

Michael Jay Mayhew
US Air Force Research Laboratory
525 Brooks Road
Rome, NY 13441
Michael.Mayhew@rl.af.mil

Michael Atighetchi
Raytheon BBN Technologies
10 Moulton Street
Cambridge, MA 02138
matighet@bbn.com

Rachel Greenstadt
Drexel University Department of Computer Science
3141 Chestnut Street
Philadelphia, PA 19104
greenie@drexel.edu

ABSTRACT

Collaborative environments, specifically those concerning information creation and exchange, increasingly demand notions of trust and accountability. In the absence of explicit authority, the quality of information is often unknown. Using Wikipedia edit sequences as a use case scenario, we detail experiments in the determination of community-based user and document trust. Our results show success in answering the first of many research questions: Provided a user's edit history, is a given edit to a document positively contributing to its content? We detail how the ability to answer this question provides a preliminary framework towards a better model for collaborative trust and discuss subsequent areas of research necessary to broaden its utility and scope.

Categories and Subject Descriptors

D.4.6 [**Operating Systems**]: [Information Flow Controls, Access Controls]; H.1.1 [**Models and Principles**]: [Value of information]

General Terms

Experimentation, Security

Keywords

Collaborative Trust, Cyber Analytics, Wikipedia

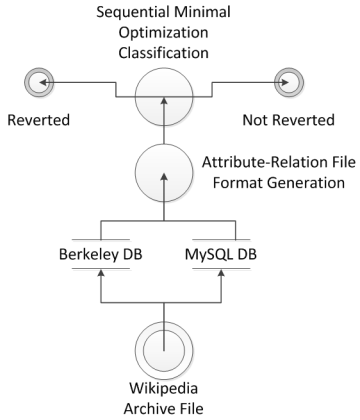

Figure 1: The data extraction and experiment flow.

1. INTRODUCTION

In today's collaborative environments, it is becoming increasingly important to associate trust values with information exchanges across a disparate set of collaborators. While enterprise service capabilities, such as Wikipedia, have started to make information available, aspects such as authority and reputation of information and trustworthiness of actors that have contributed to the creation of shared documents is often unclear and difficult to extract in meaningful ways from the underlying platform. This omission can easily lead to situations in which participants start acting on accessible, but non-authoritative, information or start collaborating with authenticated and authorized peers unaware of the fact that their behavior has recently become suspicious.

This paper describes an approach for computing trust values for information developed in environments which follow community-based workflows based on observable behavior of documents and contributing actors over time. The current work focuses on predicting whether edits to pages on Wikipedia will eventually get reverted. Until the edits are reverted, they make the document less trustworthy. Figure 1 shows a high-level view of the processing architecture and the flow of data from initial Wikipedia archive to classifica-

tion result.

The paper is organized as follows: Section 2 briefly describes related work. Section 3 describes our view on trust in collaborative environments. Section 4 describes current experimentation and results in analyzing actor behaviors in an attempt to predict future reversion of edits to pages. Section 5 concludes.

2. RELATED WORK

Wikipedia has been a good target for efforts aiming to assess interactions in a collaborative document environment in which trust plays an important role. [9] analyzes behaviors around the Wikipedia barnstars reward model. The goal of this work is to move beyond simple enumeration of a behavior (i.e., Bob has received 10 barnstars) to being able to characterize how others see behaviors (e.g., Bob likes to edit History articles and is considered helpful to others).

Geiger and Ribes [2] analyze the ecosystem of participants and bots and the process by which they coordinate to block vandalism and preserve order within the community. However, Halfaker, Kittur, and Riedl warn against overzealous blocking by providing an analysis of its detrimental effect on new participants [3].

Page et. al. [10] present PageRank, a link analysis algorithm that assigns numerical weights to a set of hyperlinked documents. Currently, the BBAC system aims to determine trustworthiness without a deep content analysis of the documents themselves.

Kamvar, Schloser, and Garcia-Molina [4] analyze reputation in peer-to-peer file sharing networks. However, the "documents" analyzed in a P2P setting are not the collaborative works seen in Wikipedia, but are instead each their own entity and anomalies are almost entirely malicious in nature.

Kittur and Kraut provide a methodology for assessing quality of artcles and assessing the relationship between the number of contributors, the coordination methods they use, and the quality of the resulting article [6]. Kittur, Su, and Chi investigated the various proposed indicators of trustworthiness and whether visualizing these features would affect users' perceptions of trust [7].

Raph Levien produced related work on trust in collaborative communities when he developed an attack resistant trust metric for the Advogato community [8].

3. COLLABORATIVE TRUST MODELS

We divide trust models into two coarse-grained categories, differentiated by the way in which roots of trust are declared.

Consensus-based models: These models represent the actions of communtities following their own specific model for assigning trust values. Wikipedia not only contains complex social structures, but those structures have evolved over time and are represented in both the behavior of long time particiapants and in the structure of the wiki itself. For example, there is a significant award culture in Wikipedia which has evolved as a way of recognizing and motivating significant contributions to the encyclopedia.

Authoritative models: Roots of trust in the authoritative model are defined by a hierarchically structured graph of authority. Representative examples of these models include organizationsl charts, access control policies, and authentication attribute services.

Calculating trustworthiness of information developed by a collaborating set of actors needs to consider the following points:

Transitivity: Trusted users produce trusted content, and trusted content is produced by trusted users. In the context of Wikipedia, assigning an award to a user is a special type of content. If the user giving the award is trusted and the award is a positive award, the recipient is more trusted. Furthermore, transitivity is declared to be imperfect. In other words, if A trusts B and B trusts C, A trusts C, but not as much as A trusts B or B trusts C. The general idea is to try to learn how much trust is "lost" through each transition.

Circularity: The two transitivity statements from a circular graph of trust relationships between actors and content. Algorithms for assigning and propagating trust values in the graph need to be able to handle circular relationships to avoid going into endless recursion.

Roots of trust: One way to break these circles is by identifying roots of trust that provide an initial anchor point of trust assignments. In Wikipedia, *Featured* and *Good* articles make for a good starting point as a root of document trust, while *Administrators* and *Bureaucrats* do the same for user trust.

This leads to the formulation of the **"Graph Problem"** for determining trustworthiness in collaborative environments. For a user U, determine the degree of trust in U based on U's contributions to content. For content, C, determine the degree of trust in C based on a) its relation to other content and b) the trustworthiness of actors that contributed to the content.

A solution to the graph problem needs to provide answers to the following decision questions:

- *Should revision R to content C be allowed?* Note that this is essentially an access control decision that needs to be performed inline with a request.
- *Should a page status be revised?* Authoritative content determination, that is the process to determine whether content is trustworty or not, is an offline decision.
- *Should a user be banned?* Examples include account suspension as a result of offline analysis.
- *Should content get locked or removed?* Examples include removal of content as a result of offline analysis.

Our attempt in solving this problem in the context of Wikipedia is to use page awards as the root of trust. Users providing revisions to trusted pages are themselves trusted. Users producing reverted revisions to trusted pages are expected to not be trusted or at least trusted less. For example, let's assume user A provided revisions to a trusted page, acquiring some level of trust. Later on, A granted an award to user B. B has a higher level of trust because of the award. The main focus is on learning what constitutes a trusted editor of a high-quality page (i.e. one with communitity consensus, namely *Featured* and *Good* pages. The specific learning context is to predict whether or not a revision to a page will be reverted.

4. ACTOR BEHAVIOR IN WIKIPEDIA

4.1 Feature Extraction

Though our research initially began with English Wikipedia, the substantial amount of data it contains became a hurdle for data manipulation. As such, we refined our data source to Simple English Wikipedia (SEW), a subset of the English Wikipedia created for those learning the English language. SEW prides itself on offering articles with a simplified grammar and vocabulary without sacrificing information quality. Just as in English Wikipedia, SEW allows, and encourages, any user to make changes to an existing article or to create new articles. The similar community structure, but significantly smaller size of SEW has allowed us to focus on the algorithms necessary for classification and learning instead of those for handling big data.

Like most MediaWiki-backed sites, archives of Simple English Wikipedia are available from WikiMedia. These archives exist in an XML format and contain the complete edit histories for most pages. Some pages, most notably those in the "Special" category, are not included with the archive as these pages have no text and are instead generated on demand. Each article revision contains a unique revision identifier, the editing user, a timestamp, and the full article text at the given snapshot. Even at the reduced size when compared to English Wikipedia, is still difficult to time consuming to use to run individual experiments. We extracted features piecewise into a relational database system and built training and test data sets directly from the database into the Atribute-Relation File Format (ARFF).

We determined the edits that were eventually reverted using a time-ordered list of page contributions along with a Murmur2 64-bit hash [1] of the page snapshot content and a content length (in case of hash collisions). For each page, two edits with the same hash and length are considered to be equivalent. Thus, edits between equivalent snapshots are no longer included in the page and can be considered reverted.

4.2 Features

Our feature set is still growing, but can be grouped into three feature categories:

- **User Features:** Information about individual users and their community status
  - user titles (administrator, bureaucrat, bot)
  - edit_count (and reverted edits)
  - revert_percentage
  - *user awards*
- **Page Features:** Information about pages and pertinent awards.
  - is_good
  - is_verygood
  - *category information*
- **Edit Sequence Features:** Information about specific page edits.
  - timestamp (including time delta since last edit)
  - length of snapshot text (including length delta)
  - is_reverted

Features in *italicized* print are not included in the experiment discussed later in this paper. These features are either currently being extracted from the Wikipedia archives and added to the database or are otherwise omitted from the current data set. Future work in this research area includes the addition of these features to training sets for use in classification models.

4.3 Classification Approach

Utilizing the ground-truth edit sequence data, we first attempt to automatically determine whether or not a new edit made to Simple English Wikipedia will eventually be reverted. Our current experiment exists as an initial benchmark of system performance and data feasibility. A core component of determining overall user and document trust is the ability to predict whether a given user is capable of successfully contributing to a document or conversation. If a user is likely to have article edits reverted, their inherent level of trust may drop along with the level of trust granted to any document containing their edits.

Classification was performed using the Weka machine learning toolkit from the University of Waikato. The toolkit provides a workspace and Java API for performing machine learning and data mining tasks and easily integrates with individual algorithm packages. Specifically, our experiments utilize sequential minimizal optimization [5] [11] library to build a model of reversion behavior and classify unknown instances.

The classification data set for this experiment included over 270,000 edit sequence instances, which come from the full edit histories of a random selection of 10,000 pages. In addition to each instance's features, as described above, each instance contains the ground-truth information of reversion as a simple true/false value. It is this value that separates the data into two classes in order to build the SVM model during training and that the classification algorithm attempts to determine during testing.

4.4 Experimental Results

Our initial results show a 97% correct classification rate of potential reverted edits to articles over the data subset. Table 1 shows the absolute and relative errors in classification. Table 2 shows the true and false positive rates as well as the F-score and ROC area for each class of behavior.

The classification model was built in approximately 2.5 hours. Of the 281539 instances used for testing, 16239 reverted edits were correctly classified along with 256849 non-reverted edits. The algorithm misclassified 1983 non-reverted instances as reverted and 6468 reverted instances as non-reverted. Pertinent are the false positive and false negative rates for the experiment as they represent anomalies in classification. Classification is done on-line and averages 1.27 microseconds per instance.

The low (0.8%) false positive rate is an encouraging starting point for system accuracy. In an online system, a false

Table 1: Classification Results Summary

Correctly Classified Instances	96.9983%
Incorrectly Classified Instances	3.0017%
Mean Absolute Error	0.03
Root Mean Squared Error	0.1733

Table 2: Detailed Classification Results

Class	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
True	0.715	0.008	0.891	0.715	0.794	0.854
False	0.992	0.285	0.975	0.992	0.984	0.854

positive represents a positive document contribution that is automatically deleted as bad information. Too high of a false positive rate may discourage users from making contributions. The false negative rate, representing the percentage of negative contributions that are permitted, is somewhat higher at almost 28.5%. A large false negative rate may lessen the integrity and overall trust in an edited document, but provides a better starting point for future editors. In other words, while 28.5% of untrustworthy edits may become part of the document, the remaining 71.5% that would have become part of the document without the classification system are automatically removed, reducing strain on editors who may remove the changes on subsequent edits.

5. CONCLUSION AND NEXT STEPS

Continuation of this research requires a fleshing out of the feature set used for classification. While current results are promising, the accuracy of the algorithm can be improved given additional data. The immediate next steps focus on higher order features relating both to user edit histories and to the category structure of articles.

The category structure of Wikipedia provides a high-level insight into how articles are related and into which articles pertain to a variety of user-defined topics. We can use this information to further our knowledge of how each individual user contributes to the encyclopedia as a whole. While we can currently determine whether a user has made a good or bad contribution to an article, we have no information relating the topic contents of a user's contributed edits. For example, a user may have a large percentage of reverted edits on one topic that, with current algorithms, overshadows a demonstratable knowledge of another topic. By analyzing not only the quantity and quality of contributions in a user's history, but also the subject areas of articles edited both well and poorly, we can better estimate a level of trust in a user's ability to successfully contribute to that field in the future.

With an improved ability to determine and recognize levels of user trust, we hope to expand into the measurement of document trustworthiness. Each article on SEW is itself a document of a given topic with a list of contributors, both good and bad, and each with their own areas of expertise. Wikipedia has procedures for the promotion and demotion of "Good Articles" and "Very Good Articles" in its own attempts at a community-driven document trust model. Though the credentials required for a page award often go beyond the combined merits of its contributors, such a model provides a ground-truth foundation for the exploration of automated trust determination.

Wikipedia is merely one platform where the notion of document trust has relevance and was chosen as a research platform due to its open nature. However, the application possibilies of our research span multiple venues, both open and closed. Government and military installations may desire a community-based method of trust for documents to either replace or enhance the traditional, hierarchical systems already in place. Businesses may see additional efficiency in an automated trust system that can decrease load on management. In academia, trust in research documents and academic papers may be derived from the trust in paper authors. The algorithms developed as part of our research stand to impact a variety of settings moving forward.

6. ACKNOWLEDGMENTS

This work was sponsored by the US Air Force Research Laboratory (AFRL). The authors would like to thank John Benner of Booz Allen Hamilton and Jonathan Webb of BBN for their research contribution.

7. REFERENCES

[1] A. Appleby. Murmurhash. https://sites.google.com/site/murmurhash.

[2] R. Geiger and D. Ribes. The work of sustaining order in wikipedia: The banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer Supported Cooperative Work*, pages 117–126, 2010.

[3] A. Halfaker, A. Kittur, and J. Riedl. Don't bite the newbies: how reverts affect the quantity and quality of wikipeida work. In *International Symposium on Wikis and Open Collaboration*, 2011.

[4] S. Kamvar, M. Schlosser, and H. Garcia-Molina. The eigentrust algorithm for reputation management in p2p networks. In *Proceedings of the Twelfth International World Wide Web Conference*, 2003.

[5] S. Keerthi, S. Shevade, C. Bhattacharyya, and K. Murthy. Improvements to platt's smo algorithm for smv classifier design. *Neural Computation*, 13(3):637–649, 2001.

[6] A. Kittur and R. Kraut. Harnessing the wisdom of crowds in wikipedia: Quality through coordination. In *Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work*, pages 37–46, 2008.

[7] A. Kittur, B. Su, and E. Chi. Can you ever trust a wiki? impacting perceived trustworthiness in wikipedia. In *Proceedings of the 2008 ACM conference on Computer Supported Cooperative Work*, pages 477–480, 2008.

[8] R. Levien. *Attack Resistant Trust Metrics*. PhD thesis, UC Berkeley, 2004. Draft Only.

[9] D. McDonald, S. Javanmard, and M. Zachry. Finding patterns in behavioral observations by automatically labeling forms of wikiwork in barnstars. In *International Symposium on Wikis and Open Collaboration*, 2011.

[10] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford InfoLab, 1999.

[11] J. C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research, April 1998.